PILOT MONITORING STUDY:

REVIEW AND FINAL RECOMMENDATIONS PREPARED FOR THE MONITORING STUDY GROUP

STATE BOARD OF FORESTRY

Dr. Don C. Erman
Professor and Director
Centers for Water and Wildland Resources
University of California
Davis, California

with assistance from:

Dr. Nancy A. Erman
Research Specialist

Ian Chan
Research Assistant

January 22, 1996

ABSTRACT

The purpose of this project is to provide an independent review of all components in the Monitoring Study Group's Pilot Monitoring Program, including both implementation and effectiveness monitoring. It evaluates the instream monitoring component completed by the California Department of Fish and Game (Rae 1995) and the hillslope monitoring component completed by CDF and Dr. Andrea Tuttle (Tuttle 1995). This report comments on the techniques that were evaluated and developed during the Pilot Program and discusses their appropriateness for the long-term monitoring program.

In summary, the review found that effectiveness monitoring approaches developed for the hillslope monitoring program are adequate. Similarly, the instream parameters tested in the pilot should provide useful information for both trend and project monitoring, but this data is not fully interpretable without similar data obtained from reference sites. For both instream and hillslope monitoring, determinations should be made *now* regarding how the data will be used to make decisions in the long-term program. If possible, hillslope and instream monitoring should be conducted together in the same watersheds. This may make it possible to link data collected on hillslopes and stream channels with multiple regression models. Reliance on demonstration watersheds, however, may not be useful in the long-term program since they may not present a realistic range of practices due to the "demonstration character" of the basins. The Pilot Program collected sufficient data to compute broad estimates of appropriate sample sizes to use in the long-term program. Finally, it would be beneficial to complete further analyses of the existing invertebrate samples collected for the Pilot Program.

The following items in Part I refer to specific requirements specified in the contract with the University of California. Part II presents in-depth reviews of: 1) the macroinvertebrate monitoring component from the instream pilot program, 2) Mr. Chris Knopp's report titled "Testing Indices of Cold Water Fish Habitat," and 3) the Hillslope Monitoring Forms developed during the Pilot Program.

PART I. CONTRACT REQUIREMENTS

3 a. Determine if the effectiveness monitoring forms for the Hillslope Component can adequately provide the types of information needed for acceptable effectiveness monitoring. Suggest changes in the forms if additions or modifications are found to be needed.

The final forms for monitoring effectiveness of application of best management practices in timber harvesting are thorough and logically complete. Dr. Tuttle has presented the options well and indicated the choices for linking the hillslope and instream programs. What is missing in this and the instream section is some rationale and discussion for how all parties will make decisions on the results. The utility of a monitoring program will be greatly enhanced by discussing standards or guidelines for making a decision from the results just as much as presenting the methods and protocols for obtaining the results. My approach follows perhaps a traditional scientific/statistical decision-making approach. The PMP team may decide on others, but to ignore this element of the program is to defer it to a time when results are gathered and then argue about their meaning-the very thing such a collaborative plan for monitoring was hoping to avoid.

I would argue strongly that the program should at least initially link monitoring for hillslope and instream components. I will give examples of important tests that result below. In response to the alternative posed in the Hillslope Report that the "demonstration watershed" approach may be best initially, I would take exception. I think it is obvious that this approach may be excellent for testing and refining methods, training personnel, and other important activities. But, it is certainly not going to present a realistic range of practice (both implementation and performance) because of its demonstration character. I believe all parties want to know whether practices are implemented and some idea of how often for what situations, whether they work when they are implemented, and what the difference in outcomes might be between results when practices are not implemented and when they are. For such determinations, there is no substitute for random selection or stratified random selection of locations for monitoring. Let's admit that to the limit of human nature, we will likely all do our best on a demonstration watershed.

If hillslope monitoring is conducted in isolation from instream, then the decision making questions relate to the following:

1. Is it clear from the existing forms how the data can be analyzed to test (qualitatively or quantitatively) the hypotheses posed by the monitoring (e.g, do practices work, how often are practices not used, what is considered "acceptable performance", etc.)?

2. Is there a process for evaluating numerical data that can inform analysts about the variation in reporting or scoring activities?

3. Given any data from the surveys, how (statistically or subjectively) would it be summarized and judged to answer hypotheses or make decisions about watershed or hillslope effects?

4. Will there be a research design element in selecting projects that can help analysts in estimating sample size, probabilities (if used in testing), or likelihood of detecting significant differences?

Specific comments regarding the Hillslope Monitoring Forms are included in Part II of this report.

---

3 b. Determine if the instream monitoring approaches provide information useful for project and trend monitoring. Determine if there are other approaches for project and/or trend monitoring which are likely to be more reliable, efficient, and effective.

The instream variables suggested and used in pilot testing should provide useful information for trend and project monitoring. However, neither monitoring results will be fully interpretable without similar data obtained from reference sites as a comparison and calibration for trends and project effects. Similar questions arise in this component as in the hillslope about decision-making from the results.

For in-depth reviews of macroinvertebrate sampling and v-star, RASI, and D-50, see Part II of this report.

There remains a troubling inconsistency in what exactly you hope these methods will reveal: troubles at a site or troubles along the way (cumulative effects). The introduction of the Instream Report seems fairly definite that the selection of these techniques and measures was based on their ability to detect cumulative effects. Yet the rationale for selection of monitoring sites (in this pilot project and presumably in the application phase as well) is hinged to exact proximate location to a RECENT, ADJACENT timber harvest. Section IV, paragraph one states "The reach commenced...where all of the direct effects of the associated timber harvest could be felt."

We have noted on several previous drafts from the MSG the conflict in purpose and methods. Presentations to the Board of Forestry on monitoring have apparently repeated the feeling that these procedures are not intended to measure effects of a timber harvest per se (thus no need for reference streams). The Instream Report continues the ambiguity of purpose and possibly of the objectives of monitoring.

The conclusion reached that excessive training is not needed is belied by the report of the process and the results. Even with trained, permanent agency scientists, and the level of review, revision,

checking, correction, resampling utilized, the error was large. Could volunteers, non-technical specialists really do better? If so, there are larger questions being raised here than implied by the study: agency professionals have trouble conducting what is thought to be routine measurement, recording, and monitoring. This result is not trivial and deserves explicit discussion and consideration for future monitoring and conclusions from it.

---

3 c. Determine the steps that would be needed to be taken to identify linkages between channel conditions and hillslope practices for the long-term monitoring program.

If hillslope and instream are linked, then how and which data from the hillslope program will be used as independent variables ("causative agents") to explain instream conditions or changes? My first guess is that some kind of multiple regression model might be used. (Stronger approaches are possible and effort should be devoted to exploring statistical or other decision-making alternatives.) Depending on the number of independent variables drawn from the hillslope program, the number of sites and streams may need to be large and sampled in the same season. Sample size might be reduced in subsequent years after testing shows that there are intercorrelated variables which can be eliminated. Regression has the advantage of forming empirical relationships that may prove useful in predicting effects (change in invertebrate response) with only measurement of the independent variables. A proposed option in the Hillslope Report, of course, is to incorporate the instream component later. I would argue the reverse on the basis that in later years variables may be dropped and predictive equations used to simplify the monitoring and the data would be at hand to support such choices. Without prior data and demonstration of relationships, the value of hillslope only data will be much reduced.

---

3 d. Determine whether and/or under what conditions trend monitoring is likely to provide useful information in managed watersheds.

The proposed methods and rationale for monitoring overall trends (changes in condition over time) in the LTMP should be adequate to detect change. There are three other major elements that require attention in order to obtain the most useful information in deciding if there has been change in response to land use activities: selection of sites, time period of observation, and reference conditions.

Selection of Sites

Because the methods used to detect change and the likely responses of the instream components to change are similar in project specific or trend monitoring, the selection of sites will be critical in separating these patterns. Obviously, if projects occur near sites chosen for monitoring long-term trends, a response of instream components may be confounded. (But Erman et al. ,1977, discussed cases in which localized disturbance could be distinguished from other project effects because of different macroinvertebrate community responses.) This fact means that selection of trend sites may not be totally random (within a basin), and that the utility of a site for trend monitoring will require some dedication and agreement with land owners on the need for and conditions of such a site.

Time

Trend monitoring means measurements over time, thus prior decisions are necessary on the length of time that trends are expected to be monitored as well as on the frequency of observation. The recent paper by Bryant (1995, Pulsed Monitoring for Watershed and Stream Restoration. Fisheries 20(11):6-13) discusses rationale and options for taking measurements in pulses when long-term monitoring is required. Because concern about land use effects in a watershed will remain as long as activities continue, the need for trend sites would remain as well. There is no established theory to predict when systems will show a response to cumulative effects, thus our present safeguard rests on observations over time both to detect response and to learn.

Reference Sites

As discussed in other sections of this report, results of instream monitoring can seldom be evaluated against absolute standards. Hence, reference conditions (streams, sites, samples) are necessary to distinguish impacts of forestry operations from natural variability or other factors. This point is well made in the Hillslope Component by Tuttle. The need exists not only to investigate the causal linkages between Rules and instream impacts, but also the need is equally strong for drawing clear inferences about results from trend (or project) monitoring. In the case of trend monitoring, a period of observation at a site becomes the "reference". In this case, the interaction of natural events (e.g., floods or "stress testing" in Tuttle's report) with watershed disturbance may show a significant departure from an established trend. But without a reference site, there will be difficulty in identifying or separating natural from human influences. Part of our work on macroinvertebrate community response to logging examined stream recovery over a 15-year period. Without reference sites, the recovery of individual streams would have been nearly impossible to detect.

---

3 e. Determine a conceptual framework for determining the best instream and effectiveness monitoring techniques to use in different types of watersheds (e.g., varying geologic conditions, resources at risk, etc.).

The current set of techniques used in the pilot monitoring program were chosen before the application to specific conditions. Some sites were eliminated based on a variety of factors, including whether the techniques could be applied. In other cases, methods were changed to improve resulting data (different amounts of subsamples counted in invertebrate samples, more intensive transect procedures for hillslope evaluation). Selection of other techniques that better fit specific circumstances requires two elements: recognition of field (or office or lab) situations where the current methods are ill-suited, and testing of alternatives. There is a tendency, especially in a regulatory framework, to seek an "approved" set of methods from which contractors, those being regulated, or the regulators themselves then feel free from further justification. Of course, established protocols in physical sciences such as water chemistry justify a formal sanctioning of accepted methods. However, in other sciences the best method is frequently a function of the site specific conditions.

For example, years of baseline monitoring of invertebrate communities by different investigators conducted in the oil shale region of Colorado showed very poor agreement even on the list of species (Erman 1981, Environmental Management 5:531-536.). One of the reasons for poor agreement was use of "standard" collecting methods (the square foot Surber sampler) in poorly suited conditions. In some larger rivers, the average rock size in sample sites was equal or greater than the size of the sampler. In the present case, the general procedures for the macroinvertebrate sampling may be well-suited to streams with the right riffle sections of appropriate size and substrate. But there will likely be situations, either coastal or Sierran, when larger streams will have cemented or armored substrates (where kick samples are ineffective) or when smaller streams have no clearly defined gravel riffles. In such circumstances the better choice than using an accepted method is to use a more relevant technique, for example, qualitative patch sampling or artificial substrates. The decision about when not to use an approved method obviously requires an understanding about stream ecology distinct from being well-trained in sessions on how to use tools. The instream component final report logically recommends pre-examination of potential sites in order to select appropriate techniques and which parameters to monitor.

Therefore, one recommendation is to devote resources during the initial phases of implementation of the monitoring program to judge whether the sites chosen fitted the techniques subsequently employed. A second, and linked recommendation is to conduct parallel studies aimed at testing the efficacy of alternative methods to classes of conditions not suited to general methods in use. Remember that the basis for judging an impact is best made against reference conditions. Using a variety of methods may weaken the capacity of the program to use all samples (sites) in broad-scale correlation analysis (as I discussed previously). But for purposes of detecting change, consistency of method between "treatment" and "control" will still provide a basis for decision.

A diagram of choices and implications is shown in Figure 1 for different conditions and sampling purposes.

---

3 f. Make recommendations concerning sample size, location, and timing of effectiveness and instream monitoring activities.

1. There must be sufficient samples of physical and biological variables within and among streams to allow for statistical tests or at least calculation of association. The pilot program has gathered probably sufficient data already to compute broad estimates of sample sizes needed to achieve various levels of precision and to examine associations between invertebrate scores and physical variables.

2. The existing samples of invertebrates should be completely analyzed. A major objective of pilot studies is to examine many of the questions I posed in my earlier review (see Part II-Macroinvertebrate Monitoring review):

a. what is the relationship between sample size and diversity (or other score) of invertebrates? The complete sample should be processed so that questions of subsample efficiency are avoided.

b. examine, if subsampling is still considered, the statistical result of sequentially increasing the size of the subsample of counts on invertebrate scores.

c. compare the cost (in time, for example) in the field and the lab for using fewer samples completely processed vs. more samples incompletely processed (i.e., subsampled) on invertebrate score precision (and accuracy, since there will be complete analysis of entire samples).

d. show by plots the change in invertebrate score with increasing sample size by randomly adding additional samples one by one. Compare the plots among streams to determine if there is a general pattern, regional differences, or individual variability.

e. compare alternative monitoring designs in terms of estimates of mean with varying confidence intervals for stated levels of accuracy (see similar comments in hillslope section).

f. use such alternatives to guide the future work in deciding how many samples and streams would be minimally necessary and optimally useful for decision making (e.g., effectiveness of practices, non-exceedance of standards).

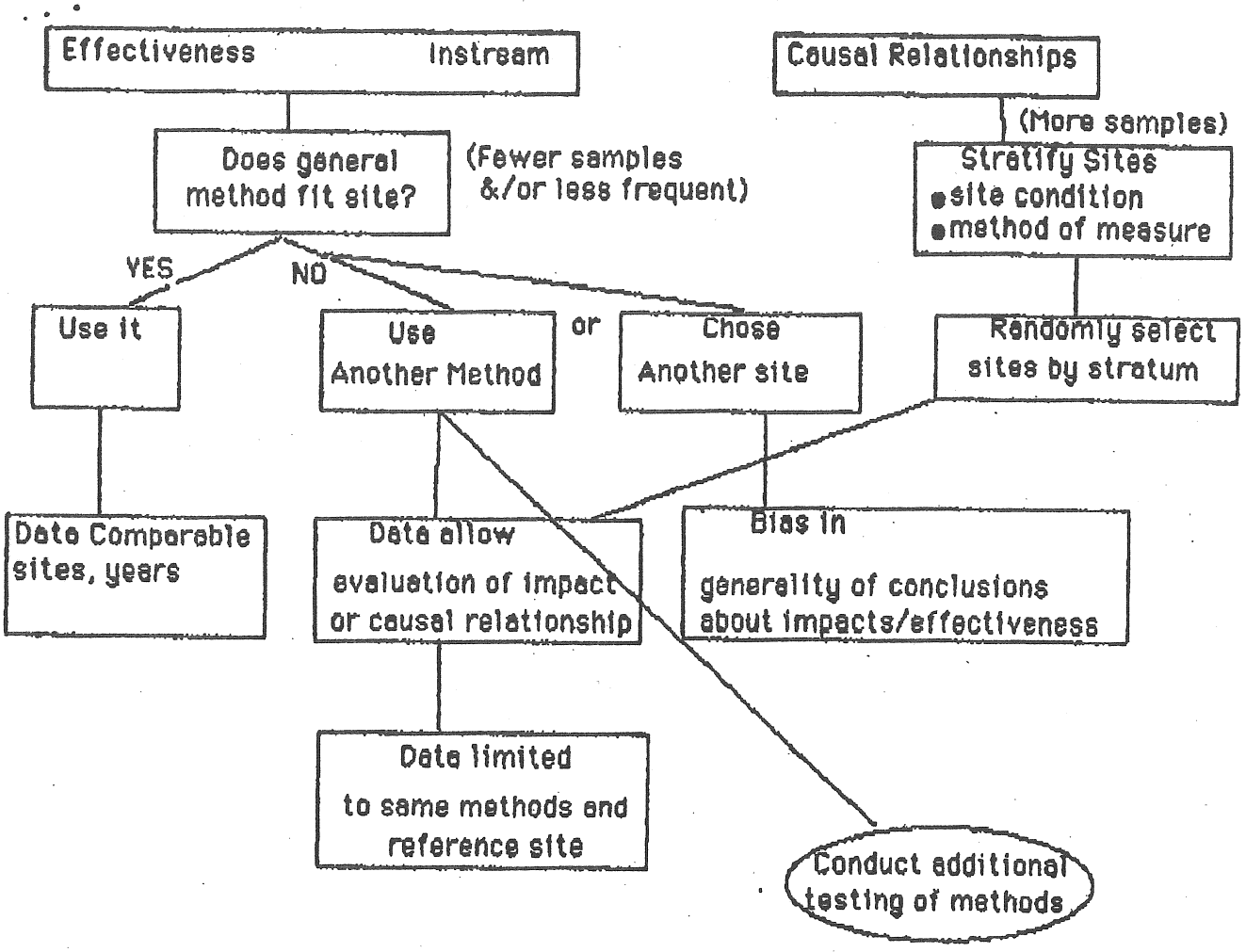

Effectiveness
Instream
Causal Relationships
(More samples)
Does general method fit site?
(Fewer samples &/or less frequent)
Stratify Sites
•site condition
•method of measure
YES
NO
Use it
Use Another Method
or
Chose Another site
Randomly select sites by stratum
Data Comparable sites, years
Data allow evaluation of impact or causal relationship
Bias in generality of conclusions about impacts/effectiveness
Data limited to same methods and reference site
Conduct additional testing of methods

Figure 1.

Before embarking on new data collection, many of the suggestions and questions presented above could be answered from the existing data collected in the pilot study. Time has no doubt limited the team's ability to fully explore many of these dimensions; take the time now before proceeding further. Will the implemented program be judged a success mearly because all parties have agreed on how to sample? Or will it be judged a failure because no one agrees on the meaning of the results? Efforts now to formalize how the data will be used to make decisions is equal in importance to how the data are to be collected. The pilot program has probably invested enough and refined the methods enough to begin the sifting and winnowing through field application. Clear ideas of how the data will be used will also benefit this refinement process.

---

3 g. Review the progress being made in producing a scientifically valid database and make recommendations for improvements if necessary.

Instream component

Data presented to date from both physical and biological samples should continue to be recorded in both raw form (unsummarized) and in conventional summary by mean, standard deviation, etc. V-star data should not be recorded beyond three decimals (perhaps two) because of the nature of the calculations and the percentage nature of the values.

Reach gradient is not only estimated at bankfull discharge and according to Dunne and Leopold (1978) gradient at this flow becomes "flattened". Gradient determined from topographic maps are partly a function of map scale, however, detailed map-derived gradients estimated in a study by Kurt Fausch (the reference was not located) compared closely with field-derived estimates. Methods for gradient measurement are not reported in the final instream component. General guidelines are that gradient is measured over a distance of at least 200 feet for a site which is within the values reported for this study.

Temperature data were measured by continuous recorders. Such voluminous data do not substitute for pre-defined criteria for judging change or significant differences. Recent papers in Water Resources Bulletin (1995) that use conventional flow-duration analysis applied to oxygen, temperature and other physical variables may provide an efficient and effective means of summarizing temperature data for assessment.

Canopy measurements are presented in the final report only as averages for each of the three methods. Standard summaries should include number of observations, means, and standard deviations. Particle size distribution data are reported only as the median particle size (D-50) although the basic data are stored for analysis by a special computer program. It would be

useful to compute and report also other percentiles of the distribution to convey the spread of the distribution. Conventional percentiles in addition to the 50th are 25 and 75 and/or 16 and 84.

Hillslope component

Several elements of the hillslope component await further testing. Included in this category is the sensitivity of the data collection process to differences in the evaluators. This work should continue (as suggested in the final report) both as a means of evaluating monitoring responses and as an additional step in judging the need for training. The records from the data forms for hillslope evaluation are extensive. I have commented earlier that the MSG might give some thought to optical scanning forms or other methods to assist in efficiently summarizing data from each THP.

The final report discusses difficulties in evaluating effectiveness with implementation in one of the four possible cases (implementation is only evaluated when problems are identified). A partial solution was to add a second evaluation for a transect as a whole. An alternative (and complimentary procedure) would be to stratify THPs and on a random subset of evaluations conduct a more rigorous analysis of this case with an additional form. This procedure would allow the MSG or others to examine the assumption that a "transect as a whole" judgment provides adequate information with reasonable accuracy and is highly recommended.

The qualitative evaluation of the entire THP is a legitimate process for rating performance (conditions) when conducted by trained professionals. This rating process complements the more structured, quantitative transect approach. If the subjective rating procedure is accurate then there are statistical procedures for testing ranks of summary scores from the transect samples against the qualitative ranks. The hypothesis is obviously that ranks from both procedures are the same.

In this component, as in the instream, the MSG should summarize the data already on hand as an exercise in exploring how judgments would be made based on the data. The purpose of such an exercise is not to draw conclusions about the Rules at this stage, but rather to verify that: 1) collected data will answer the questions, and 2) all data collected are necessary.

---

3 h. Determine how the on-going monitoring program could be supplemented by or integrated with future research programs, including work undertaken at the University of California's Wildland Resources Center.

This report raises a number of situations in which further testing is needed. Well designed experiments or structured observations would greatly facilitate decisions about adequacy, alternatives, and assumptions inherent in the monitoring program. A standard practice should be if a major assumption is necessary to interpret results, then a test of the assumption is probably worth doing. The UC Wildland Resources Center can be a vehicle for involving the academic community in many of the future testing that has been suggested. There is, for example, work underway to evaluate sampling procedures for small streams lacking defined riffles.

The Center can also simplify contract arrangements when a variety of UC researchers may be involved on different aspects of the overall program. For example, the traditional agency approach is usually to contract individually with each investigator for each subproject. An alternative that the Center can provide is exemplified by the Sierra Nevada Ecosystem Project (SNEP). The overall project is managed by the Center and subcontracts to consultants or university researchers are specified by the overall project needs. Only one contract to the University is required. Selection of specific subcontractors is made through project needs and proposal solicitation. For a possible program with CDF, the selection process also would have an advisory panel (that included non-University personnel) that would evaluate proposals and potential contractors.

The Center also maintains an up-to-date directory of wildland expertise throughout the UC system. The Center has regularly in the past made contact with relevant researchers when projects are contemplated, when outside agencies are looking for expertise, and for other purposes. Queries of experts and faculty with relevant interests could be done as a means of seeking collaborative projects with CDF or others in the MSG.

---

4. Prioritize the recommended monitoring steps in light of the high probability of constraints in funding.

1. The MSG should complete various analyses with data collected during the initial phase of testing and method development. Detailed examples for the instream component have been previously described. Although the purpose of this phase of the project was not to evaluate THPs or test effectiveness, the data set are of the type that the actual monitoring program will obtain. Treating the existing data as an example will help spot troubles now and alert the MSG to potential difficulties. Analysis at this level may also permit estimations of sample size necessary for statistical testing. The most important point of such an exercise would be to simulate how all parties would use the data to reach decisions about whether the Rules are followed, are adequate, and similar questions.

2. As I discussed previously, some agreement prior to actual data collection should be made now on how the data will be used to make decisions. This process may include choices of statistical significance level, what constitutes adequate implementation, or what constitutes a "failure" in performance that monitoring is designed to detect.

3. Complete field testing of remaining pieces of the overall program, such as the Whole THP hillslope evaluation process.

4. Initiate simultaneous hillslope and instream components in a random (or stratified random) selection of sites in the three regions.

5. Begin field scoping and owner discussions to locate long-term (trend) sites.

6. Begin ancillary projects to develop alternative and/or complimentary methods, test assumptions, compare methods.

7. Review results from first year of application of program.

PART II: IN-DEPTH REVIEWS OF INSTREAM AND HILLSLOPE MONITORING TECHNIQUES

MACROINVERTEBRATE MONITORING USED IN THE INSTREAM COMPONENT OF THE PILOT MONITORING PROGRAM: Biological Assessment of forested streams using benthic macroinvertebrates.

The broad goal of this portion of the assessment ("...to assess the effectiveness of field analytic techniques to detect significant changes in water quality on private timberlands") is neither achieved nor necessary to test. The use of macroinvertebrate communities as a valid assessment of logging impacts has been established for some time and has been demonstrated specifically for logging in California (Erman, et al. 1977, Erman and Mahoney 1984, and others). The current study is a modification of traditional stream invertebrate impact assessment only in its use of a random subsampling of individual organisms collected in samples and a summation of several indices of potential disturbance as suggested by the EPA and others. To that extent, the study objectives are trivial.

In earlier discussions with members of the Monitoring Study Group, considerable emphasis was placed on the need to simplify the methods used to monitor and to avoid "research scale" techniques that were assumed to be too expensive in time and money for routine monitoring of timber harvest and best management practices. We are pleased that, except for the continued omission of reference sites in the study design, the proposed monitoring of invertebrate communities is for all intents and purposes similar to that used in research: many uniform, randomized, and consistent samples (we assume) are taken and organisms are counted and identified to the lowest practicable unit. Because the project was assumed, however, to be significantly new and untested (either as a general technique or as an application in California) it has failed to take advantage of existing knowledge and has added little, not even another test of the objective of assessing the effectiveness of the proposed methods. Closer adoption of standard methods used by others previously would improve data quality, comparability, and efficiency than the proposed process. A few comparisons with previous work will illustrate these points.

The project relied on a D-shaped net to collect organisms disturbed from a fixed area on the stream bottom The dimensions of the net must be specified so one is able to judge whether it can quantitatively or qualitatively collect the organisms carried downstream from the sample area. Since the area disturbed is reported to be 1 ft x 2 ft, the method effectively collects the equivalent of two square foot Surber samples. The Surber sampler is designed to collect uniformly and with confidence from gravel riffles of small streams by its construction and instructions for use. The sampling method used in this study may be a satisfactory alternative to a conventional Surber type sampler if used with consistency, but the general strictures similar to a Surber sampler must be followed for achieving quality results.

The mesh size of the D-net specified in this project is 0.8 mm, a size numerous studies have shown is too large for quantitative and in some cases qualitative collection of small organisms. For example, the list of organisms from Table 5a lists Copepoda and Ostracoda. These taxa are too small for inclusion in this study because of the large mesh size. The mesh size also calls into question the counts of individuals (and hence taxa number and dependent indices) comprising Chironomidae and Hydracarina. Presence of these taxa in the samples should not be taken as evidence of sampler effectiveness. Small macroinvertebrates can appear in collections of fish made with a minnow seine. Either use a smaller mesh size or discard those taxa unlikely to be retained effectively in the 0.8 mm mesh.

The micro-habitats in streams are numerous. Riffle sampling by use of Surber-like techniques is a constraint of the tool but the general advice for studies aimed at making comparisons among streams is to keep the sampled micro-habitats as similar as possible. For example, it is customary to restrict samples to areas of similar depth, velocity, and substrate or to incorporate the habitat variation in a stratified design to avoid bias. The present study controlled for position (edge to edge and stream center) and to some extent substrate (well-graded substrate) but purposely included an uncontrolled manner/ leaf packs, aquatic vegetation, and woody debris as sample points. If the purpose were to compare micro-habitat richness among streams such a procedure might be acceptable for the goals of this study, their inclusion adds additional complexity and could inflate taxa richness in those streams so sampled. In addition, the study did not specify how many samples (of the 9 on each stream) were allocated to these different habitats. Obviously, small differences in the number of samples by habitat type, even if two streams had the same variety of micro-habitats, could skew resulting metrics. The study data are not presented in a way that the importance of this sampling difference can be examined. At a minimum, such sample allocation must be reported unless all streams are treated the same.

In explaining the methods for the study, one should define the methods measured. For example, the study refers (p. 9) to a diversity index such as in Shannon and Weaver (emphasis mine), and a "Modified" Hilsenhoff Biotic Index. No definition is given as to the actual metrics used. Was the diversity index the Shannon index and how was the Hilsenhoff Index modified?

The total amount of stream bottom sampled from each stream was 18 ft^2 (nine collections each containing 2 ft^2). Previous work on the use of macroinvertebrate communities to detect effects of logging (Erman et al. 1977) has shown that eight Surber samples are sufficient. For greater precision in estimating abundance and diversity, Erman and Mahoney (1984) presented a method for increasing sample size while reducing processing (this paper is cited in the references for the instream study). They showed that four Surber equivalents from four clusters of four Surber samples were adequate and reduced further processing in the laboratory to only four Surber samples per stream. The present study has chosen instead to take an exceptionally large number of samples from each stream and then to rely on partial counts of

the material to reduce effort. This procedure leads to major bias in some of the resulting metrics.

Random subsampling for counting and identification have been demonstrated effective when the objective is to produce certain indices. Much work has been devoted to an examination of this topic, especially as it relates to the Sequential Comparison Index. This work found that a minimum of 200 counts of individuals was usually required for consistent values of the Index. The present study repeats in part this known relationship but errors in the interpretation of the change in going from counts of 100 to 300 individuals. For the indices reported (Diversity Index and Biotic Index) the relationship between the values based on 100 or 300 individuals is highly associated linearly ($r^2 = 0.929$ for Diversity, $r^2 = 0.843$ for Biotic Index). However, by increasing the count to 300 individuals, in several cases all or nearly all of the organisms in a sample are counted and identified. When this result occurs (some samples are completely processed, others are not) other metrics are disproportionately affected. For example, in the 1993 collections where all organisms in a sample were processed, the percentage increase in the number of taxa decreased exponentially with the percentage of the total individuals represented by a count of 100 (Figure 1, log transformation of both variables, $r^2 = 0.914$). Because taxa richness is often related to total number of individuals examined in a sample or to the size of the sample, taxa richness has lost its value as an unbiased variable in this procedure and reflects primarily the percentage of the sample counted rather than the sample or the response to possible land use effects. A similar weakness (bias in estimating taxa richness) will also affect the values in the EPT Index, which is no more than the sum of the taxa in three orders.

The use (and abuse) of functional feeding groups has no stated purpose in this study and its connection to land use is unclear. The categories are based on an extreme generalization for the taxa and not intended for site-specific use without detailed examination of actual food habits. No use is made of them in the study.

Overall then, the proposed assessment cannot use density (counts of individuals are fixed), number of taxa is biased, EPT is biased by taxa bias, and according to the report, the dominant taxa metric is too variable and the biotic index is based on an untested and inappropriate relationship (organic pollution). Therefore, of the proposed metrics in the present study, only the index of diversity remains as a reliable dependent variable, a fact well established by previous work. Obviously, a solution to this part of the problem is to analyze the entire sample and take fewer samples, i.e., adopt the established methods used by previous research. More consistent data with less effort (both in the field and in the lab) will result.

The combined index, as shown on p.32, which sums the ranks for several metrics, is flawed because of the points made above. In any event, although the stated purpose of the study was to detect differences in water quality, there was no association with the qualitative physical/habitat scores (no data presented, however), and no use of "the more quantitative

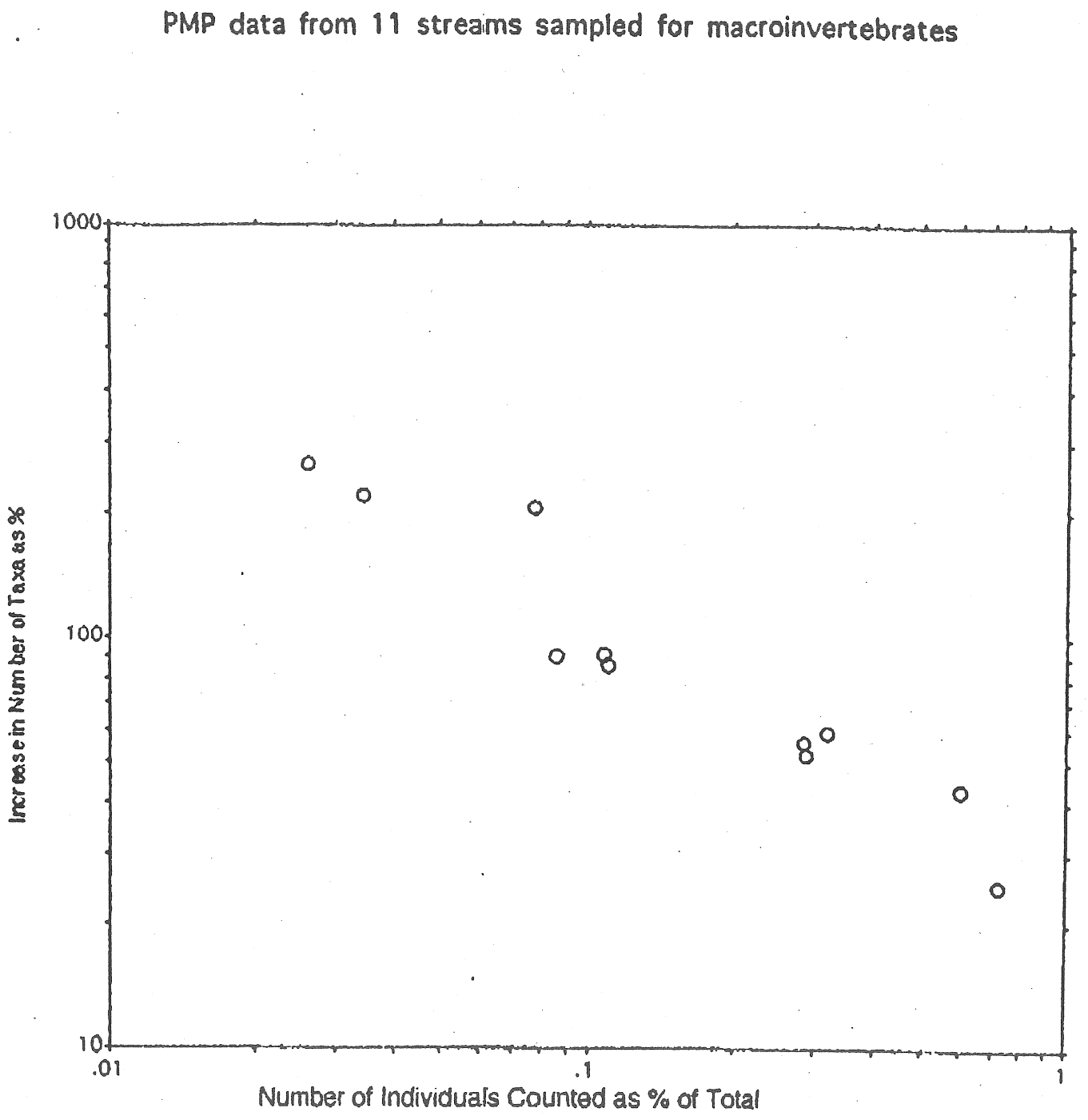

PMP data from 11 streams sampled for macroinvertebrates
1000
100
10
Increase in Number of Taxa as %
.01
.1
1
Number of Individuals Counted as % of Total

Figure 1.

measurements collected by the PMS crews." Given the stated goals of this project, a complete data analysis should be provided as recommended in the Conclusion.

The study has used the coefficient of variation (CV) as a measure of suitability of the variables to detect land use effects. Values of CV are reported and judged adequate except for the Dominant Taxa metric. No reference to other work on the variability of these measures is presented. Again, better use of the substantial literature in this field would be useful both in judging the adequacy of the sample size and in comparing this study to others for relevant metrics. For example, Roby et al. (1978, Freshwater Biology 8:1-8) in a comparison of Surber samples with artificial substrates found that CVs for 8 Surber sampled ranged from 18.3% to 27.6% (mean 19.8%) in estimates of Shannon diversity. Erman and Mahoney (1984, Water Resources Center Contribution 186) showed that the procedure of cluster sampling discussed above resulted in CVs for Shannon diversity that ranged from 0.9 - 3.0% and for taxa richness that ranged from 6.6 to 10.1%. These values compare to those in present study of 2 - 9% for diversity (Shannon?) and 5 - 17% for taxa richness on subsamples of 300 individuals. Mahoney and Erman (1984) further illustrated the relationship of sequentially adding additional Surber samples to the change in Shannon diversity. They showed, as others have, that beyond about 4 Surber samples, there is little change in the estimate of diversity. The present study evaluated the change in metrics from counting individual subsample and the total sample. From this comparison, 300 individuals were selected but no data or basis was given for this choice. A better approach commonly used is to compare the effect of 100, 200, 300, etc. and total count (as an X axis) on the response of various metrics (Y axis) so that a logical basis for subsample size or diversity can be determined. However, the problem is best avoided entirely by processing an entire sample, as discussed above.

Nevertheless, except for the problems of the cost of large sample size (18 ft^2 per stream reach), lack of use of a vast literature on this subject, testing procedures which are standard and already proven, and creating bias from the subsampling of the total sample, the basic approach follows common practice and with proper execution should yield valid results. The problem which is unresolved in this entire study is how any of the results can or will be interpreted when monitoring begins.

COMMENTS ON "TESTING INDICES OF COLD WATER FISH HABITAT" BY MR. CHRIS KNOPP

The study "Testing Indices of Cold Water Fish Habitat", which was an early product of the general effort to improve monitoring in California, is a valuable contribution to this goal. A subsequent comment letter, solicited by the California timber industry, found fault with the study and raised several criticisms. I provide here some of my assessments of this study.

1. There is an emphasis in the project placed not so much on the evaluation of "which physical elements of instream habitat are affected by human activity in the upslope watershed" as the evaluation of two relatively new response variables to watershed activity: V-star and RASI. As a consequence, even though the design was rigorous (the number of test streams was large for field-based tests of hypotheses), there will remain in the eyes of some scientists a level of skepticism about the relationships and utility of these newer variables. Generally, monitoring employs well-established, time proven methods. In the present study, for example, the only source cited for RASI is to Kappesser (1992) in an unpublished Forest Service report. Such lack of standard peer-review weakens acceptability of a technique to be used routinely.

2. The presentation of the data has emphasized the testing of the hypotheses about upslope disturbance in relation to the variables measured to detect disturbance. Although this approach is fully appropriate for the main purposes of the study, I recommend additional data exploration (and have conducted some here) from the substantial data set obtained. Categorization of streams into levels of disturbance for purposes of the primary analysis created non-continuous groups (e.g. low, medium, high). But the actual response variables (RASI, V-star) should fit a continuous distribution; that is, all the streams should fit a single relationship. The report conducts some exploration (admittedly limited by resources to sample and stratify additional sites) of interactions among the variables which brings forth several additional insights. Examples are the display of pattern of response of landuse levels versus stream slope and drainage area. The means for addressing these sources of interaction with the main variable was to eliminate some sites and "match" the categories of land-use level by similar stream gradients or basin area. Again, in the context of the analysis, this approach is appropriate; but for wider understanding of relationships, other approaches might also be useful. For example, the covariates of slope and area could be included in an Analysis of Covariance to account for any significant trends due to those variables prior to testing for differences among management level. Or subsequent to ANOVA one could use multiple or several simple regressions among variables for all sites. When I conducted such regression analyses it showed that the inclusion of a few sites with large drainage areas accounted for most of the bias (see Figure 1) in the overall relationship of, for example, V-star. A similar conclusion was reached in the study; however, simple plots illustrate the relationship without omitting sites.

There is value in using all the available data in various multiple regressions with the dependent variable alternately V-star, RASI, D-50. Some variables were not estimated for all sites and

Knopp data plots

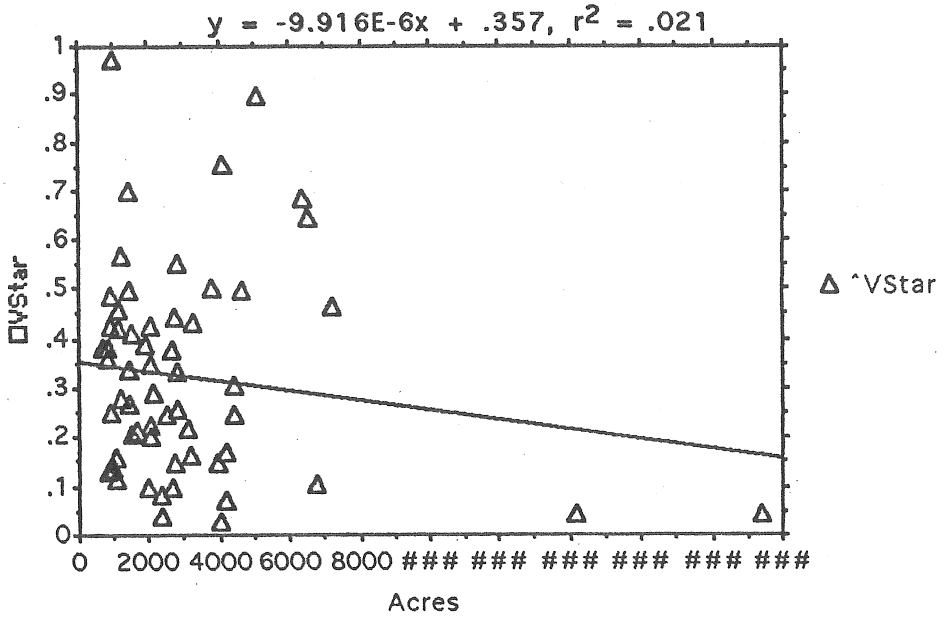


FIGURE 1

hence the number of variables that can be included in regression was limited. As an example, in regressions with five independent variables (Exponential Recovery, V-star Pool Volume, Log_{10} (Wood Volume), Log_{10} (Acres), Reach Slope) it appeared that the best predictors of RASI were exponential recovery, log (wood volume), and log (acres). Although all three variables had highly significant partial regression coefficients, the entire model (all five independent variables) only explained about 30% of the variation in RASI.

One of the components of V-star is V-star pool volume. This variable is highly correlated with V-star ($r^2 = 0.707$, Figure 2), and I am unsure if this variable is volume with or without fine sediment. One should ask, however, whether the pool volume variable itself was sufficient without further computation of V-star. Detecting the "true" bottom of a pool partially filled with fine sediment is likely to be difficult to standardize. Pools go through episodes of scouring and filling during flood flows (Bjornn et al. 1977. Bull. 17, Forest, Wildlife, and Range Exp. Sta. Univ. Idaho; Andrews, 1979. USGS Prof. Pap.1117) and generally contain some fine sediment from natural processes. Because the pressure applied to a rod may vary by individual investigator, uniformly probing to a "true" pool bottom is crucial but subjective. In lake benthic work, similar "penetrability" indices have been used but have relied on a free-falling rod from a fixed distance in order to minimize the problem of unequal force. Such an index, of course, does not attempt to measure the total accumulated sediment depth but development of the tests recognized the need to standardize the procedure. If the same person always applied the probe to pool bottoms, one would obtain consistency. Such an outcome may help explain the reliable results of Lisle and Hilton, who developed the technique.

3. The study concluded, through the use of discriminant analysis, that V-star and RASI were the best variables for identifying affected or unaffected streams. However, V-star and RASI are significantly associated ($p < .001$), even though the correlation is low (Figure 3, Table 1). And both variables are significantly associated with D-50 ($p < 0.001$), again with coefficients of determination low (about 30% to 60%). As perhaps expected, RASI and D-50 are virtually the same variable and have an r^2 of 60% (See Figure 4). If these two variables are basically measuring the same thing, one should ask whether RASI offers other advantages. The report suggests that it is less affected by dissimilarities in hydraulic properties than D-50; however, as this study evolved, streams with broadly dissimilar hydraulic properties (slope, drainage area) had to be excluded anyway. The difficulties attendant to estimating the largest mobile particle size (extremely subjective) and the judgmental requirements of how to select the proper particles may introduce difficulties in standardizing the technique compared to pebble counts (D-50). (For example, a recent paper (1994) in the Wat. Resources Bulletin found pebble count data also useful as a method to estimate percent fines and showed short- and long-term recovery from disturbance.)

4. The report makes an assumption about the meaning of RASI and V-star. It implies (in the title) that these variables are fish habitat, a point criticized in the CFA commissioned review of the report. There is some justification for such criticism. These variables are hydrologic. Being new measures of hydrologic conditions they lack studies linking them to fish populations (or

Knopp data plots

FIGURE 2

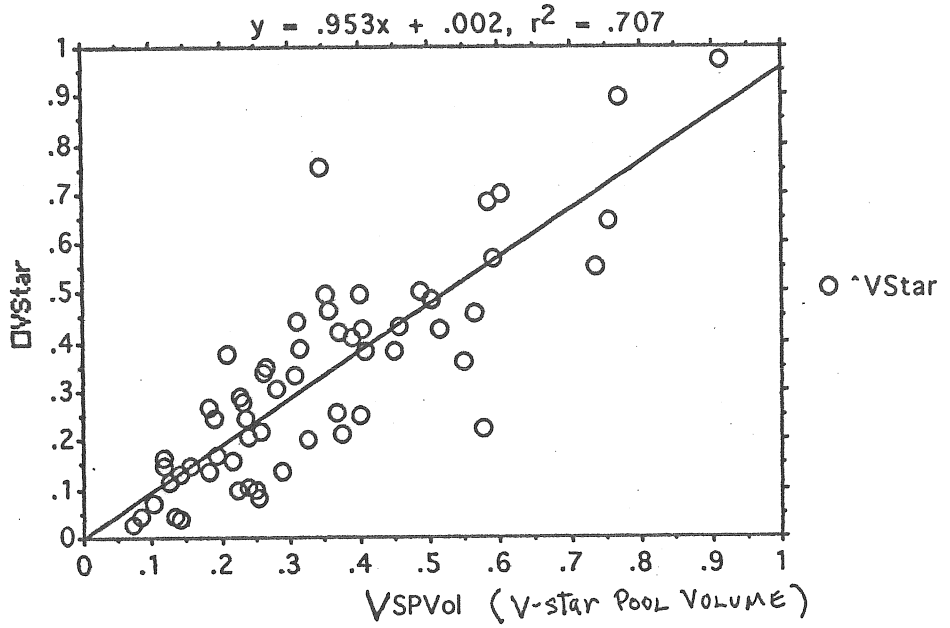
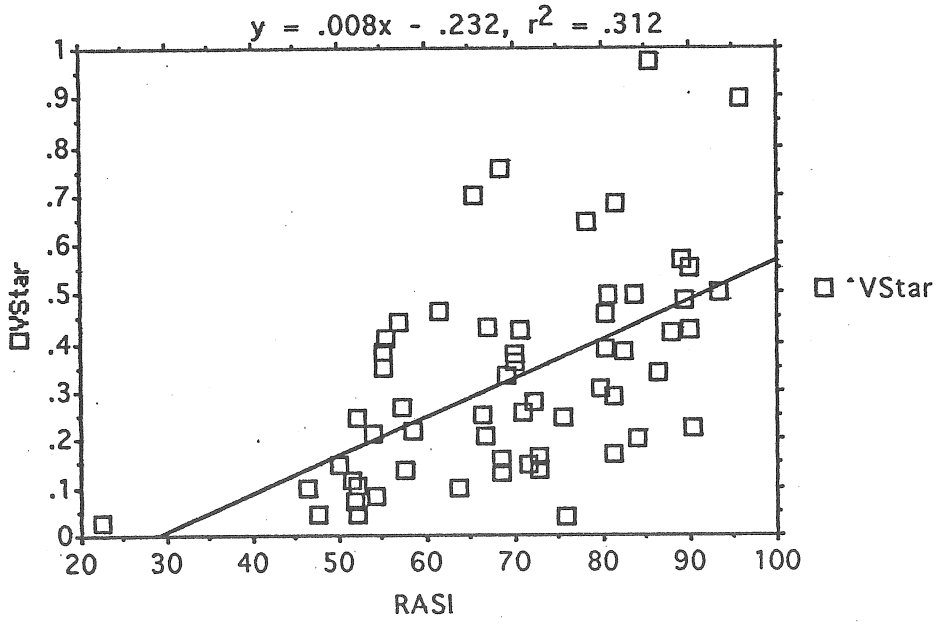


FIGURE 3



Knopp data plots

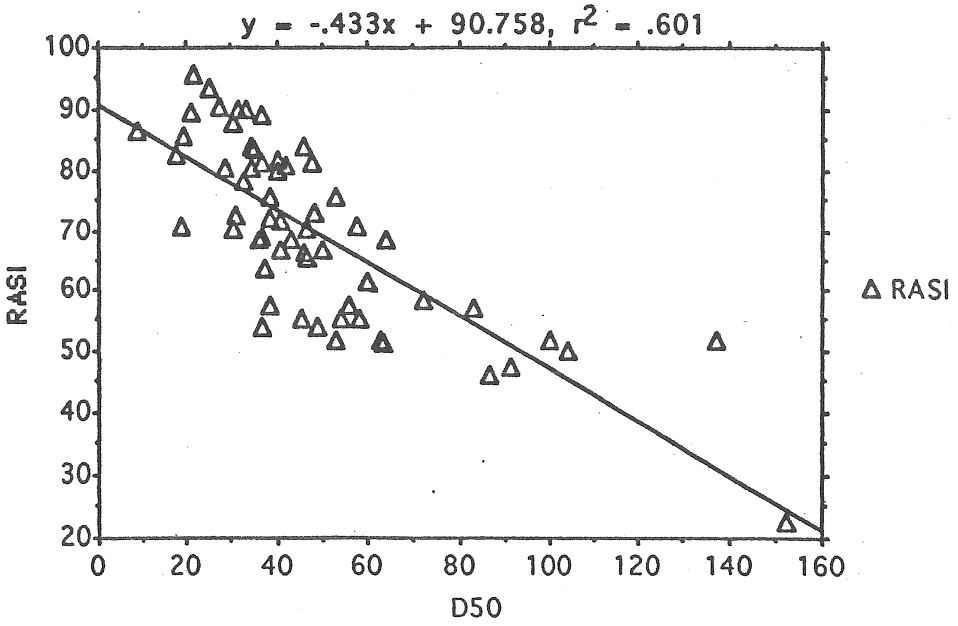


FIGURE 4

Knopp data plots

TABLE I,

Correlation Matrix for Variables: X_1 ... X_9

	SPVol	D50	RASI	^VStar	WdVol	WdCov	WdSubst	Acres
SPVol	1							
RasilD50	-.477	1						
RASI	.519	-.713	1					
^VStar	.829	-.412	.462	1				
WdVol	-.139	.273	.011	-.024	1			
WdCov	.04	-.106	.204	-.04	-.083	1		
WdSubst	-.125	.057	-.036	-.167	.036	.413	1	
Acres	-.095	.424	-.146	-.062	.697	-.134	-.163	1
RcSlope	-.231	.162	-.166	-.27	-.18	.068	-.13	.067

Note: 9 cases deleted with missing values.

any biotic response). However, the association between fish populations and pools in streams is well established, although in specific cases a species or life stage may not be linked to pool habitat. Nevertheless, the report is a thorough evaluation of some hydrologic variables that may reflect land use activities in a basin. RASI, V-star, and D-50 all relate to various attributes of stream sedimentation. All three discriminated in both conventional parametric and non-parametric tests among streams with different land use intensity. There should be no dispute about the problem of accelerated erosion rates and stream sedimentation, quite apart from the issue of whether or not these variables reflect specific fish habitat.

5. The volume of large wood, wood cover and other associated attributes of wood in the stream are known to be important for stream hydraulics, sediment routing, and biotic habitat as pointed out in the study. Unfortunately, the practice, until recently, of actively removing large wood from streams, (and continuing even now in some streams) especially in North Coast streams, makes current measurements of these variables of unknown significance to understanding past or current land use activities. A lack of significant correlation (although the sign is in the correct direction, negative) between wood volume and stream gradient for the streams in this study also suggests an atypical response.

COMMENTS ON THE HILLSLOPE MONITORING FORMS FOR THE RANDOM TRANSECT APPROACH GENERATED FROM THE PILOT MONITORING PROGRAM

1. The evaluation as a whole must be more consistent in the structure of the forms and the reference to the parts. The term transect form is used, for example, when the form is the Effectiveness Rating (I think). Never refer to the same thing with more than one term. Some Effectiveness Rating forms are not transect based. The Effectiveness forms are each set out in a different format. One is not sure in some cases which is the actual form (WLPZ forms for Effectiveness are not labeled that way anywhere). The Effectiveness Rating form for WLPZ has a heading requiring "Observer", "Date", "THP No." and "Plot No." but this heading is missing for Watercourse Crossings. Such inconsistency creates considerable confusion and frustration. By contrast, the Implementation forms are much more uniform, but even here, note the inconsistency in heading and subheading styles among the forms .

2. The evaluators are asked to record all features on the transect sheet but are asked only to give an overall summary rank for non-problem features in the implementation forms. Would ranking the non-problem features require too much time? There does not seem to be a direct one-to-one connection in some of the form sets between effectiveness items that are ranked or evaluated and the list of specific rules that are ranked for implementation. Thus, many of the rules may not get ranked except in the overall column C. Is it possible that had an evaluator ranked each feature, there would have been some ranked as minor departure or even major departure without necessarily a corresponding "problem" noted in effectiveness? We think that rankings of all rules would be informative, especially if a specific rule does not have a clear effectiveness counterpart.

3. How will these data be summarized? The forms constitute a long and relatively complex survey. The form sets for effectiveness are different for each type of activity (partially a consistency problem as noted above). Some forms have items with three possible choices (ranks), others have only two choices (e.g., diversion potential in Watercourse Crossings forms). Are the various choices to be given a number rating? We presume that summary data would then be able to give averages for each item. The procedure for Part III, implementation, is to give values directly for each problem (column B) and for each transect as a whole. No simple method of averaging will be possible for Part III. Rating codes 1 and 2 may not apply for column B and a rating of 0 (cannot determine) would lower average ratings even though a problem (hence, at least a rating of 3) was recorded. Similar problems apply to rank 5 but perhaps in the other direction. A simple tally of the number of each rank will be possible and obviously, a computer code can be made to average whatever portion of the ranks are desired. If there is no intent to use the numbers as ranks, merely as codes, then we suggest referring to these numbers consistently as codes, not as ranks.

Column C presents other problems. Instructions ask the evaluator to "rate the implementation of all the rules based on your judgment of the transect as a whole." In the case of column B, the number of points is identified (thus giving both frequency and intensity of ranks), but in column C

only the judgment of overall rank is given and the total number of points is tallied in the effectiveness form (p. 3, instructions). We suggest that the evaluator at least keep score, in the same way as for problems points, the rank of "non-problem points" used in making the overall rank. The evaluator is already asked to record the occurrence of all pertinent features along a transect (instructions, p. 3). Without some background ranks in this column, understanding the meaning of column C will be difficult and impossible for possible follow-up analysis. For example, what was the proportion of minor departure vs. major departure vs. couldn't determine? Was an overall rank of 3 given because in some cases a rank of 1 was earned but they were "averaged out" by ranks of 4? And if the objective is to make an assessment of overall implementation, wouldn't you want to have each "point" or feature ranked, regardless of whether or not it was a problem? Perhaps this part is a trade-off for the time needed to fill out the forms but it seemed to me a ranking of each feature on a transect would be more straight-forward (although more time consuming).

4. Where are evaluators supposed to explain Implementation ranks 5 (and maybe NA)? At the end, under "No rule to express problem?"

5. Have you considered adapting these forms to an optical scanning form (fill in the bubble with #2 pencil) or something similar? Someone will have a tedious, potentially error-prone job of counting up the ranks by hand.

6. Numbered items in the Implementation forms are stated as questions. They could as well be given as statements. And for each statement, a corresponding rule number is given. Is this number necessary for the evaluation? We presume that the statement is a factual or close approximation of the rule. Is it needed by the evaluator to interpret the statement or give a ranking? By listing the rule number, is there possibility of bias in the evaluator: i.e., would an evaluator make a different choice if only the words of the rule were given? It may be a subtle difference, but we would recommend that the evaluators base their judgments only on the words that are given, not on what might be in their minds about their interpretation of each rule.

7. The logic of the evaluation is that each element (roads and trails, landings, etc.) has its own form set for effectiveness and implementation. This procedure makes some sense but it leads to a lot of what seems to us like duplication of rule evaluation. There must be a flipping back and forth between sets as the evaluator ranks a rule first under roads (for example) then under crossings. Technically, we understand that the official rules are numbered and come under each particular activity in the timber harvesting process. But, the actual site-how a road performs at a watercourse crossing, for example-dictates how a place is ranked and once should do it. Trials with the forms may sort out this issue.

8. Have you considered yet what would constitute "good" scores and "bad" scores (or passing?) for effectiveness and implementation?